

Análise de Dados com Base em Processamento Massivo em Paralelo

Aula 3: Processo de ETL/ELT

Cristina Dutra de Aguiar
ICMC/USP
cdac@icmc.usp.br

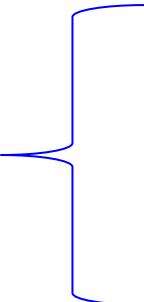


CEPID - Centro de Ciências
Matemáticas Aplicadas à Indústria

Agenda

- Contextualização
- Operações
- Exemplo usando Pandas

Processo de ETL/ELT

- Representa a atividade mais complexa, cara e demorada do *data warehousing*
 - 75% do custo
- Operações
 - Extração (Extract)
 - Transformação (Transform)
 - Carga (Load)
 - Tradução
 - Limpeza
 - Integração

Projeto do Data Warehousing

- Deve ser realizado **antes** do processo de ETL/ELT
 - Indispensável para um bom desenvolvimento do *data warehousing*
- Aspectos a serem considerados
 - **Objetivo** da aplicação de *data warehousing*
 - **Recursos** disponíveis
 - **Hardware** e **software** apropriados
 - Pessoas envolvidas
 - Planejamento da **capacidade** do ambiente

Principais Atividades Envolvidas

- Identificar o **propósito** da aplicação de *data warehousing* e o **volume** de dados manipulado
- Identificar a **arquitetura** de *data warehousing* e seus **componentes**
- **Instanciar** a arquitetura, por meio da integração de servidores, ferramentas e tecnologias
- Realizar o **projeto** do *data warehouse*, dos *data marts* e do *data lake*
- Identificar as **fontes** de dados que possuem dados relevantes
- **Integrar** as fontes de dados ao *data warehousing*

LGPD (Lei Geral de Proteção de Dados)

- Características
 - Introduz diversas **mudanças jurídicas**
 - Proteção de dados pessoais
 - Responsabilidade civil dos responsáveis pelo **tratamento dos dados**
- Possui **grande** impacto no *data warehousing*
 - Coleta, processamento, armazenamento, extração, utilização, modificação, ...

Base das Explicações

- Carga dos dados no *data warehouse*
 - Operações podem ser aplicadas aos *data marts* e ao *data lake*, respeitando-se as particularidades de cada um
- Processo de ETL/ELT como um todo
 - Não considera a abordagem na qual projeta-se apenas as operações do processo de EL para a carga dos dados no *data lake* para pré-exploração dos dados
- Manipulação de grandes volumes de dados
 - Não discute especificamente *big data* e *data streaming*

Processo de ELT e Big Data

- Introduz complexidade adicional
 - A **quantidade** de fontes de dados é muito maior
 - A variedade de **domínios** é muito maior
 - Muitas fontes de dados são **dinâmicas**
 - As fontes de dados são **extremamente heterogêneas** com relação ao formato dos dados
 - Os dados apresentam grande **variabilidade**, dificultando a identificação de mesmas entidades do mundo real presentes em diferentes fontes
 - Os dados das fontes de dados apresentam muita variação de **qualidade**
 - O **tratamento incipiente** do aspecto temporal é muito mais emergente

Diferença entre Instância e Esquema

- Instância
 - Coleção de dados armazenados no banco de dados em um determinado momento, ou seja, são os **dados** propriamente ditos
 - Sinônimos: **extensão** do banco de dados, **linhas** (ou **tuplas**) de tabelas relacionais e **registros** de arquivos
- Esquema
 - **Projeto** do banco de dados, incluindo as entidades e os relacionamentos entre essas entidades
 - Sinônimos: **intenção** do banco de dados, **definição de tabelas** relacionais e **definição da estrutura** (campos) dos registros de arquivos

Aplicação de Data Warehousing da BI Solutions



- Propósito

Foco: **salário** e
quantidadeLançamentos

Perspectivas: funcionário
cargo
filial
data

- Volume de dados

- Pequeno
- Foco em **funcionário**

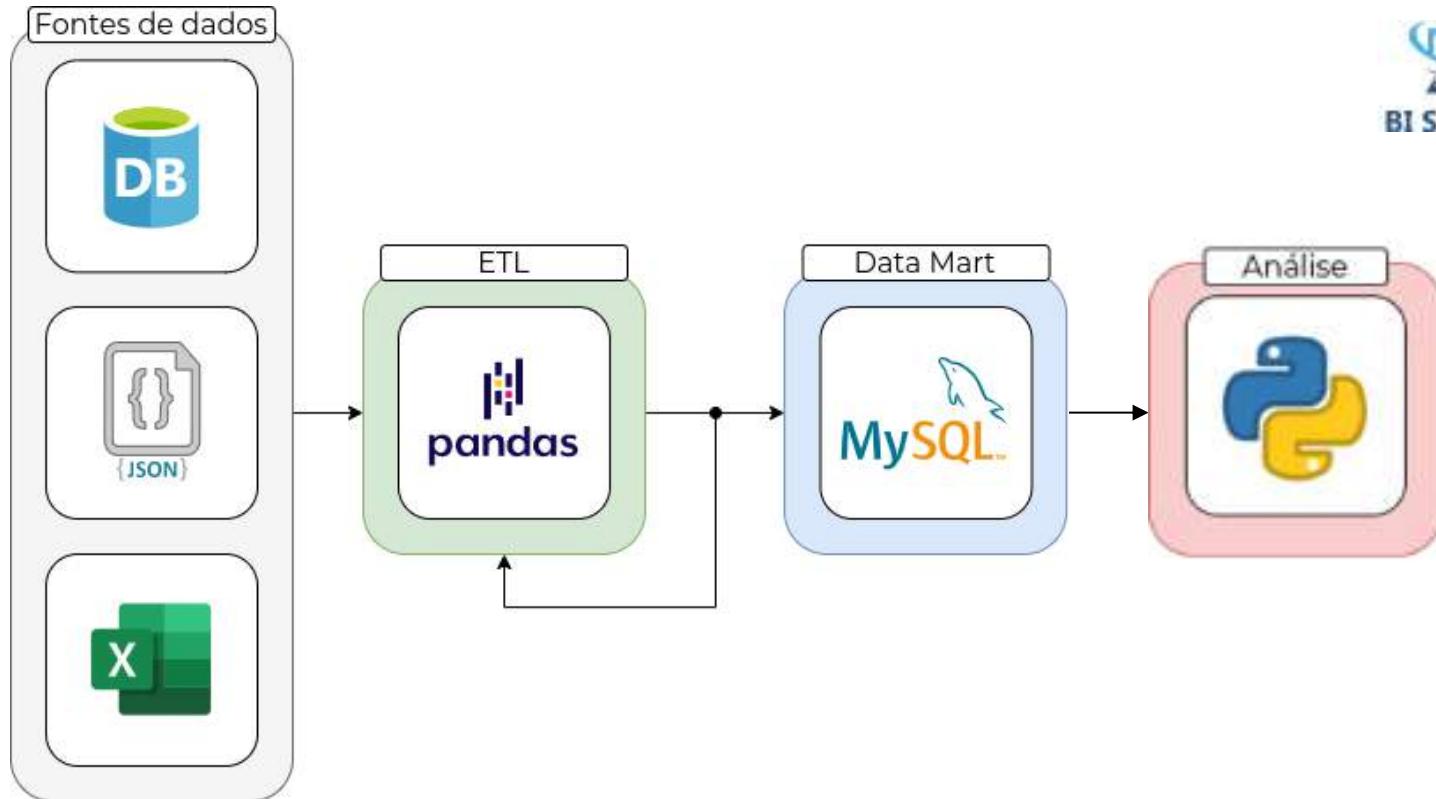
- Arquitetura

- Processamento em **lote**
- Uso de **Data Mart**
- Baseado do **modelo relacional**

- Fontes de dados

- Banco de dados **relacional**
- Arquivo **JSON** (NoSQL)
- Planilha **Excel**

Instanciação da Arquitetura: Pipeline



Projeto do Data Mart: Funcionário



Relação funcionário

funcionario (funcPK, funcMatricula, funcNome, funcSexo, funcDataNascimento,
funcDiaNascimento, funcMesNascimento, funcAnoNascimento,
funcCidade, funcEstadoNome, funcEstadoSigla, funcRegiaoNome,
funcRegiaoSigla, funcPaisNome, funcPaisSigla)

Tabela relacional

funcPK	funcMatricula	funcNome	funcSexo	funcDataNascimento	funcDiaNascimento	funcMesNascimento	...
1	40	Abdiel Lima	M	9/4/1990	09	04	...
2	1	Aline Almeida	F	1/1/1990	01	01	...
...

} esquema
} instância

Fontes de Dados

funcionário (funcMatricula, funcNome, funcSexo, funcDataNasc, funcCidade, funcEstado, funcPaís)

(Fonte 1)
Tabela
Relacional

funcMatricula	funcNome	funcSexo	funcDataNasc	funcCidade	funcEstado	funcPaís
3	ARON ANDRADE	M	03/03/90	Santos	SP	Brasil
10	ADETE CARVALHO	M	10/10/90	Osasco	SP	Brasil
...

} esquema
} instância

(Fonte 2)
Arquivo
JSON

```
[{"colab_matricula":71,"colab_nome":"Ademil Castro","colab_sexo":0,  
"colab_data_nasc":"1970-11-10",  
"colab_cidade":"Belo Horizonte","colab_estado":"MG","colab_pais":"BRA"},  
 {"colab_matricula":72,"colab_nome":"Ademor Costa","colab_sexo":0,"colab_data_nasc":"1971-12-11",  
"colab_cidade":"Araguari","colab_estado":"MG","colab_pais":"BRA"}, ...]
```

(Fonte 3)
Planilha
Excel

	A	B	C	D	E	F
1	Matrícula do Empregado	Nome do Empregado	Sexo do Empregado	Data de Nascimento	Cidade de Residência	Estado de Residência
2						
3	78	Garcia, Adenir	Masculino	17/06/77	Morretes	Parana
4	79	Gomes, Adenil	Masculino	18/07/78	Recife	Pernambuco
5

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Fontes de Dados

(Fonte 1)
Tabela
Relacional

funcMatricula	funcNome	funcSexo	funcDataNasc	funcCidade	funcEstado	funcPais
3	ARON ANDRADE	M	03/03/90	Santos	SP	Brasil
10	ADETE CARVALHO	M	10/10/90	Osasco	SP	Brasil
...

(Fonte 2)
Arquivo
JSON

```
[{"colab_matricula":71,"colab_nome":"Ademil Castro","colab_sexo":0,  
"colab_data_nasc":"1970-11-10",  
"colab_cidade":"Belo Horizonte","colab_estado":"MG","colab_pais":"BRA"},  
 {"colab_matricula":72,"colab_nome":"Ademor Costa","colab_sexo":0,"colab_data_nasc":"1971-12-11",  
"colab_cidade":"Araguari","colab_estado":"MG","colab_pais":"BRA"}, ...]
```

esquema

instância

(Fonte 3)
Planilha
Excel

	A	B	C	D	E	F
1	Matrícula do Empregado	Nome do Empregado	Sexo do Empregado	Data de Nascimento	Cidade de Residência	Estado de Residência
2						
3	78	Garcia, Adenir	Masculino	17/06/77	Morretes	Parana
4	79	Gomes, Adenil	Masculino	18/07/78	Recife	Pernambuco
5

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Fontes de Dados

(Fonte 1)
Tabela
Relacional

funcMatricula	funcNome	funcSexo	funcDataNasc	funcCidade	funcEstado	funcPais
3	ARON ANDRADE	M	03/03/90	Santos	SP	Brasil
10	ADETE CARVALHO	M	10/10/90	Osasco	SP	Brasil
...

(Fonte 2)
Arquivo
JSON

```
[{"colab_matricula":71,"colab_nome":"Ademil Castro","colab_sexo":0,  
"colab_data_nasc":"1970-11-10",  
"colab_cidade":"Belo Horizonte","colab_estado":"MG","colab_pais":"BRA"},  
 {"colab_matricula":72,"colab_nome":"Ademor Costa","colab_sexo":0,"colab_data_nasc":"1971-12-11",  
"colab_cidade":"Araguari","colab_estado":"MG","colab_pais":"BRA"}, ...]
```

(Fonte 3)
Planilha
Excel

	A	B	C	D	E	F
1	Matrícula do Empregado	Nome do Empregado	Sexo do Empregado	Data de Nascimento	Cidade de Residência	Estado de Residência
2	78	Garcia, Adenir	Masculino	17/06/77	Morretes	Parana
3	79	Gomes, Adenil	Masculino	18/07/78	Recife	Pernambuco
4
5

} esquema
} instância

Agenda

- Contextualização
- Operações
- Exemplo usando Pandas

Operações

- Extração
- Tradução
- Limpeza
- Integração
- Carga

Extração

- Objetivo
 - Extração dos dados de interesse das fontes
 - Encaminhamento dos dados para as demais operações
- Tarefas
 - Quais dados são extraídos de quais fontes de dados
 - Como esses dados são extraídos
 - Com qual frequência esses dados devem ser periodicamente extraídos
 - Qual técnica empregar para identificar dados das fontes que foram alterados

Tipos de Extração

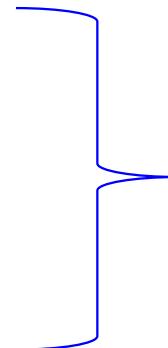
- Extração inicial
 - Carga inicial dos dados no DW
 - Quanto maior o volume de dados, mais tempo é consumido
- Demais extrações
 - Ocorrem devido às alterações nos dados das fontes
 - Técnicas que podem ser aplicadas
 - Extração completa: todos os dados são extraídos e carregados novamente
 - Extração incremental: apenas os dados que sofreram alteração são extraídos e utilizados para determinar o novo conteúdo do DW

Autonomia das Fontes de Dados

- Intervenção mínima
 - Aplicações devem prover seus dados com a **menor intervenção** possível
- Operação não intrusiva
 - Extração **não** pode **impactar negativamente** na execução das aplicações
 - **Janela de manutenção** para processamento em lote
 - Período no qual os sistemas operacionais geralmente ficam mais ociosos
 - **Extração contínua** para processamento de *data streaming*

Exemplos de Abordagens de Extração

- Processo de extração
 - cliente: aplicação que realiza a extração dos dados
 - servidor de dados: fonte de dados
- Abordagens
 - Interface de comandos padronizados
 - Protocolo comum de acesso aos dados
 - Conversor de comandos



podem ser utilizadas conjuntamente ou separadamente

Interface de Comandos Padronizados

- Interface do cliente e do servidor como um elemento comum
 - Interface genérica, ou seja, *Application Programming Interface* ([API](#))
- Aplicação identifica qual a fonte de dados e um **driver** específico **converte** os formatos e comandos
- Exemplos
 - ODBC (*Open Database Connectivity*)
 - DBE (*Borland Database Engine*)
 - JDBC (*Java Database Connectivity*)

Código da aplicação cliente **não precisa ser alterado** quando esta for redirecionada para outro tipo de servidor

Protocolo Comum de Acesso aos Dados

- Protocolo bem definido para conectar aplicações clientes aos vários tipos de servidores
 - Interface **cliente**: recebe requisições de aplicações clientes traduz no protocolo comum
 - Interface **servidora**: identifica e processa as requisições traduz os resultados para o protocolo comum
- Exemplos
 - DRDA (*Distributed Relational Database Architecture*)
 - RDA (*Remote Data Access*)
 - REST (*Representational State Transfer*)

Garante
interoperabilidade entre
tipos de servidores
mesmo que diferentes
APIs tenham sido
usadas

Conversor de Comandos

- Converte comandos de manipulação de dados e de formato de dados (*gateway*)
- Provê a funcionalidade de tradutor
 - Realiza o **mapeamento** de dados e de comandos entre os vários clientes e servidores
- Exemplos
 - *Database Gateway* para DB2
 - *Informix Enterprise Gateway*

Usualmente **complementa** ou **estende** as outras abordagens

Detecção e Propagação de Alterações

- Rotinas
 - Monitorar as **modificações** ocorridas nas fontes de dados
 - **Identificar** quais modificações ocorreram em quais dados
 - Extrair somente os **dados necessários**
- **Frequência** (periodicidade, latência)
 - Depende das necessidades das análises e do nível de consistência desejado
 - Exemplos: assim que o dado é gerado, a cada hora, diária, semanal

Técnicas Empregadas

- Chamadas de *Change Data Capture (CDC)*
- Dependem das facilidades oferecidas pelas fontes de dados
- Abordagens
 - *Timestamp* (marcadores de tempo)
 - *Triggers* (gatilhos)
 - *Logs*
 - *Snapshots* (instantâneos)

Timestamp

- Armazenado em uma **coluna de auditoria**
- CDC
 - **Compara** o *timestamp* com a data e o horário da extração mais recente
 - Extrai dados que possuem **data de alteração maior** do que a **data dessa extração**
- Exemplo
 - Kafka Connect (Confluent)
 - Somente **poucos** dados operacionais possuem *timestamp*
 - Não identifica dados **removidos**
 - Pode ser **intrusiva**

Triggers

- Presentes em sistemas gerenciadores de banco de dados relacionais
 - CDC
 - Usa triggers para a detecção
 - Realiza a notificação automática de alterações
 - Comando CREATE TRIGGER
 - Oracle
 - PostgreSQL
 - DB2
- Somente poucas fontes oferecem recursos de gatilhos
 - Podem ser intrusivos
 - Podem onerar o servidor de dados

Logs

- Armazenam **todas as transações** que ocorrem na aplicação
 - Inserções, remoções e atualizações
 - Consultas
- CDC
 - **Percorre** o arquivo de *log*
 - Identifica as **diferenças** que devem ser extraídas
- Exemplo
 - Logstash (Elastic)

- Requerem **privilégio** de administrador de banco de dados
- Possuem **formatos proprietários**
- **Protegidos** pelo sistema

Snapshots

- Foto dos valores de dados armazenados em um certo momento no banco de dados
- CDC
 - Compara o *snapshot* da extração anterior e com o *snapshot* da extração atual
 - Gera um *arquivo delta* com as atualizações
- Comparação de *snapshots*
 - Solução comumente usada

Comparações cada vez maiores precisam ser realizadas à medida que o volume de dados da fonte cresce

Projeto do Data Mart: **funcionario**



funcionario

funcionario (funcPK, funcMatricula, funcNome, funcSexo, funcDataNascimento,
funcDiaNascimento, funcMesNascimento, funcAnoNascimento,
funcCidade, funcEstadoNome, funcEstadoSigla, funcRegiaoNome,
funcRegiaoSigla, funcPaisNome, funcPaisSigla)

Tabela relacional

funcPK	funcMatricula	funcNome	funcSexo	funcDataNascimento	funcDiaNascimento	funcMesNascimento	...
--------	---------------	----------	----------	--------------------	-------------------	-------------------	-----

Extração para o Exemplo da BI Solutions

- Quais dados são extraídos de quais fontes de dados?
 - Carga inicial: todos os dados de interesse



Tabela **funcionarioRelacional**
Relacional

funcMatricula funcNome funcSexo funcDataNasc funcCidade funcEstado funcPais

colaboradorJSON

Arquivo
JSON

```
[{"colab_matricula": " ", "colab_nome": " ", "colab_sexo": , "colab_data_nasc": " ",  
"colab_cidade": " ", "colab_estado": " ", "colab_pais": " "}, ...]
```

empregadoPlanilha

Planilha
Excel

	A	B	C	D	E	F
1	Matrícula do Empregado	Nome do Empregado	Sexo do Empregado	Data de Nascimento	Cidade de Residência	Estado de Residência
2						

Extração para o Exemplo da BI Solutions

- Como os dados de interesse são extraídos?
 - Por meio de APIs
- Com qual frequência os dados de interesse são extraídos?
 - Processamento: em lote
 - Frequência da extração incremental: mensal



Extração para o Exemplo da BI Solutions

- Qual **técnica** empregar para identificar dados das fontes que foram alterados?
 - `funcionarioRelacional`: *Triggers*
 - `colaboradorJSON`: Comparação de *Snapshots*
 - `empregadoPlanilha`: Comparação de *Snapshots*



Operações

- Extração
- Tradução
- Limpeza
- Integração
- Carga

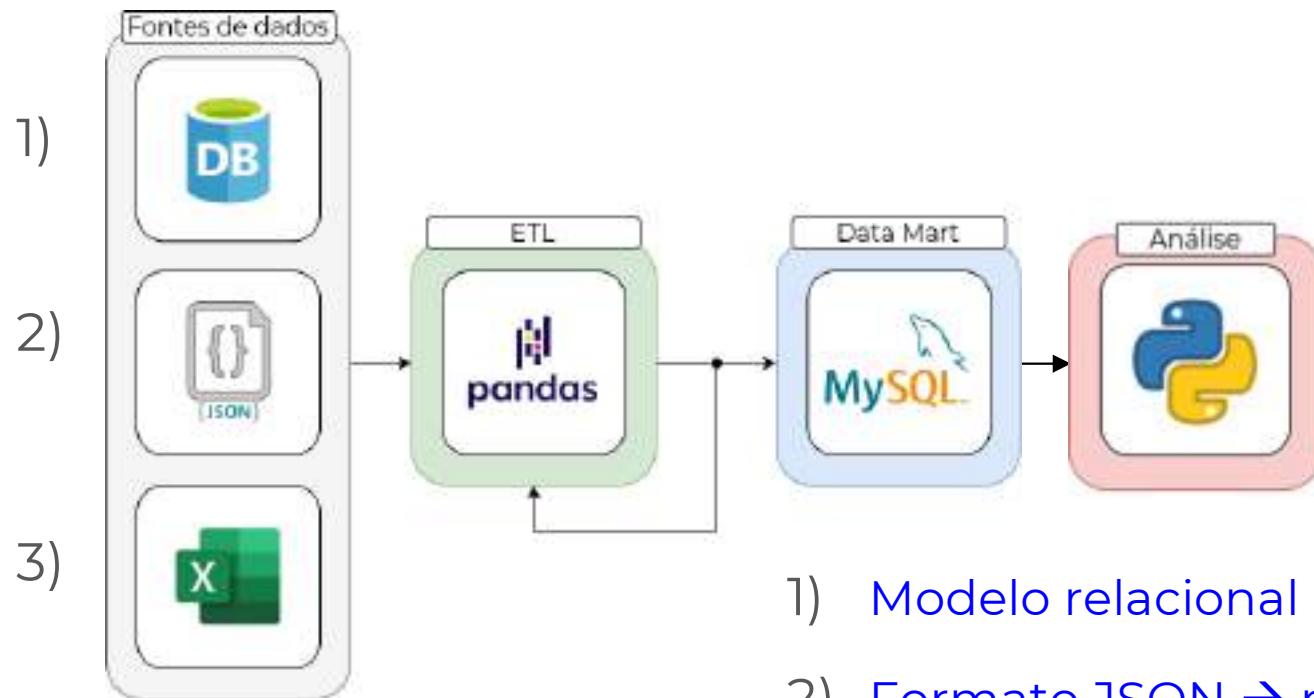
Problema

- Dados armazenados nas fontes de dados
 - Heterogêneos
 - Seguem diferentes modelos de dados
 - São representados por conceitos diferentes
 - Possuem diferentes formatos
 - Redundantes, inconsistentes e até mesmo complementares
- Dados armazenados no *data warehouse*
 - Devem seguir um projeto e uma forma de organização específica

Objetivos

- Realizar a **conversão** entre o formato nativo das fontes de dados e o formato do *data warehouse*
 - Esquema
 - Instância: valores dos dados e tipos de dados
- Garantir a manutenção da temporalidade
 - Maioria das fontes de dados **não é histórica**, mas o *data warehouse* sempre deve armazenar dados históricos
 - Dados temporais podem ser adicionados indicando o **momento de atualização dos dados nas fontes** de dados ou o **momento de armazenamento no data warehouse**

Rotinas para o Exemplo da BI Solutions



- 1) Modelo relacional → modelo relacional
- 2) Formato JSON → modelo relacional
- 3) Formato de planilha Excel → modelo relacional

Operações

- Extração
- Tradução
- Limpeza
- Integração
- Carga

Limpeza

- Garante a **acurácia** e a **qualidade** dos dados
 - Dados devem atender às restrições de integridade impostas pelas regras de negócio

Deve ser realizada durante todas as atividades do processo de ETL/ELT

- Exemplos

- Comprimentos de campos inválidos e uso de caracteres inválidos
- Dados incompletos, em branco, ou usando abreviações não padronizadas
- Duplicações dos mesmos dados (ou seja, redundância)
- Descrições inconsistentes, violação de restrições de integridade, associação de valores inconsistentes

Limpeza para o Exemplo da BI Solutions

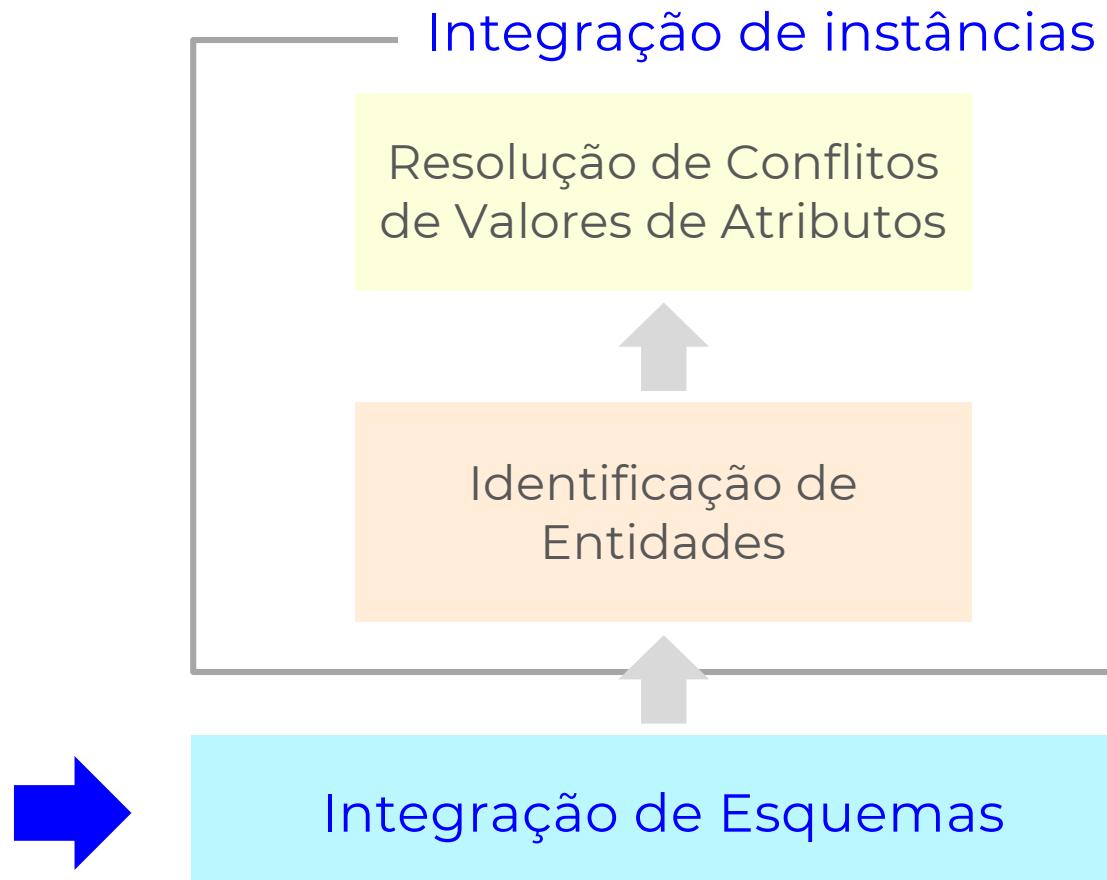
- Uso de estratégias aprendidas em disciplinas anteriores
 - Técnicas Avançadas de Captura e Tratamento de Dados
- Exemplos
 - Redundância
 - Detecção de *outliers*
 - Tratamento de informações errôneas
 - Manutenção da acurácia do **número de matrícula** dos funcionários



Operações

- Extração
- Tradução
- Limpeza
- Integração
- Carga

Visão Geral



Integração de Esquemas

- Definição
 - Especificação de **mapeamentos** que descrevem os relacionamentos semânticos entre os esquemas das fontes de dados e o esquema do data warehouse
- Relativismo semântico
 - **Conflitos** existentes entre duas ou mais representações
 - **Diferentes usuários** modelam o mesmo pedaço do mundo real de **diferentes formas**, de acordo com as suas percepções

Conflito

- Surge quando duas ou mais representações do mesmo conceito **não são idênticas**
 - Usam construtores diferentes
 - Aplicam diferentes restrições de integridade
- Tipos
 - **de nome**
 - **semântico**
 - **estrutural**

discrepâncias existentes entre os esquemas
apresentam mais do que um tipo de conflito

Conflito de Nome

- Relacionado aos **nomes** que representam os diferentes elementos nos esquemas a serem integrados
- **Sinônimos**
 - Diferentes nomes são aplicados ao mesmo elemento
 - `funcionarioRelacional`, `colaboradorJSON` e `empregadoPlanilha` indicam funcionários
- **Homônimos**
 - Mesmo nome é aplicado a diferentes elementos
 - Data representa `data de contratação` em um esquema e `data de aniversário` em outro



Conflito Semântico

- Mesmo elemento é modelado em diferentes esquemas, porém representando conjuntos que se sobrepõem
- Exemplo
 - **funcionarioRelacional**: funcionários da área de *Engenharia*
 - **colaboradorJSON**: funcionários da área de *Marketing*
 - **empregadoPlanilha**: funcionários da área de *Recursos Humanos*
 - funcionários podem ser diferentes entre si
 - o mesmo funcionário pode estar em mais do que uma fonte de dados, desde que ele mudou de área de atuação durante a sua trajetória



Conflito Estrutural

- Diferentes **construtores estruturais** são utilizados para modelar o mesmo conceito representado em diferentes fontes de dados
- Exemplo
 - **Endereço**
 - **Atributo** do esquema **funcionário** em uma fonte de dados
 - Esquema **endereço** composto de outros atributos: **nome da rua**, **número** e **complemento**



Mapeamentos (Data Mart e Fonte Relacional)

■ ■ funcionario ≡ funcionarioRelacional

funcMatricula = funcMatricula

funcNome = funcNome

funcSexo = funcSexo

■ funcDataNascimento = funcDataNasc

funcCidade = funcCidade

■ funcEstadoSigla = funcEstado

funcPais = funcPais

■ conflito de nome
(sinônimos)

■ conflito semântico

Mapeamentos (Data Mart e Fonte JSON)



■ ■ funcionario ≡ colaboradorJSON

- funcMatricula = colab_matricula
- funcNome = colab_nome
- funcSexo = colab_sexo
- funcDataNascimento = colab_data_nasc
- funcCidade = colab_cidade
- funcEstadoSigla = colab_estado
- funcPais = colab_pais

■ conflito de nome
(sinônimos)

■ conflito semântico

Mapeamentos (Data Mart e Fonte Planilha)



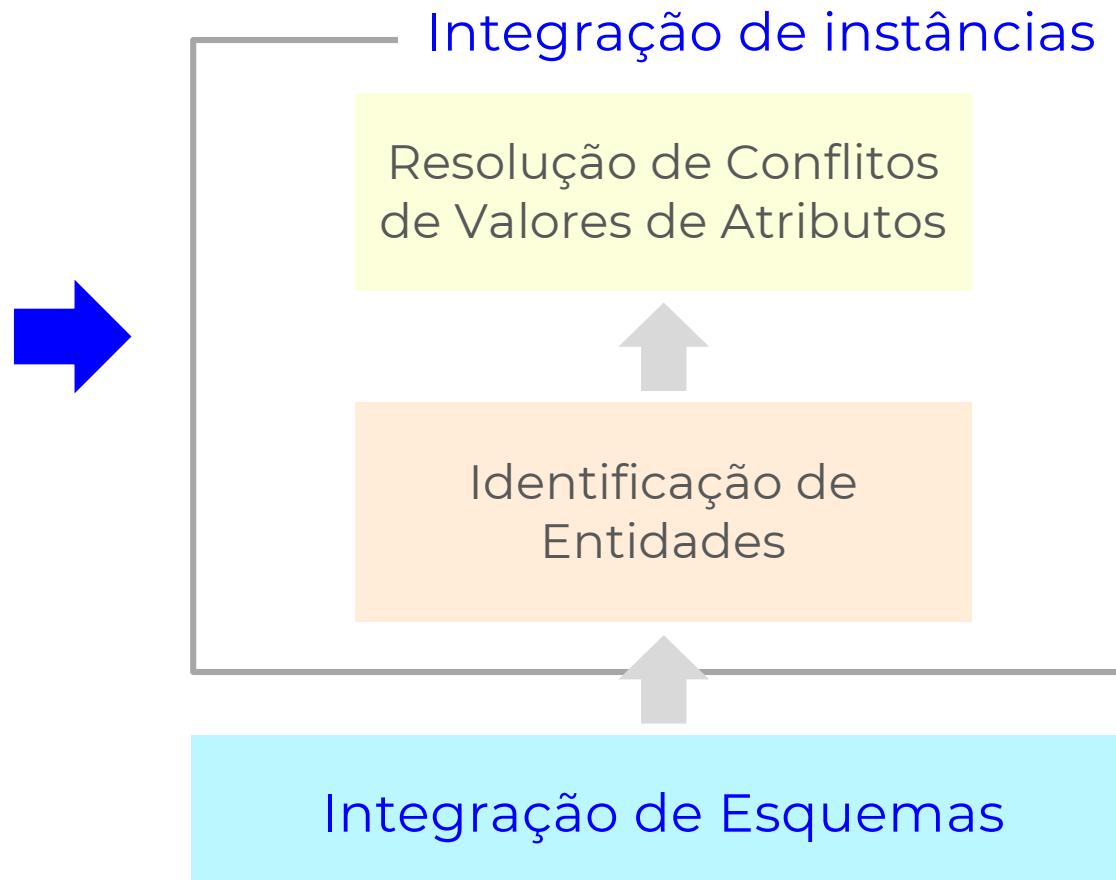
■ ■ funcionario ≡ empregadoPlanilha

- funcMatricula = Matrícula do Empregado
- funcNome = Nome do Empregado
- funcSexo = Sexo do Empregado
- funcDataNascimento = Data de Nascimento
- funcCidade = Cidade de Residência
- funcEstadoNome = Estado de Residência

■ conflito de nome
(sinônimos)

■ conflito semântico

Visão Geral



Identificação de Entidades

- Objetivos
 - Identificar quais entidades das fontes de dados heterogêneas referem-se à mesma entidade do mundo real
 - Agrupar essas entidades em agrupamentos de entidades similares
- Cenários
 - Identificação **única** das entidades por meio de atributos chave
 - **Não existe um atributo** que identifica univocamente cada entidade

Identificação Unívoca das Entidades



- Fontes de dados
 - `funcionarioRelacional`: atributo `funcMatricula`
 - `colaboradorJSON`: atributo `colab_matricula`
 - `empregadoPlanilha`: atributo `Matrícula do Empregado`
- Valores dos atributos referentes à matrícula dos funcionários
 - Analisados na operação de limpeza dos dados
 - Representam valores acurados

Resolução de Conflitos de Valores de Atributos

- Resolve **inconsistências** nos **valores** dos dados das entidades que referem-se à mesma entidade do mundo real, mas que **diferem** nos valores dos seus atributos
- Exemplos



Sexo do Funcionário

F/M

Feminino/Masculino

0/1

Nome do Funcionário

Adenildo Campos

Campos, Adenildo

A. Campos

Operações

- Extração
- Tradução
- Limpeza
- Integração
- Carga

Funcionalidades

- Realizar processamentos adicionais
 - Geração de agregações (visões materializadas)
 - Necessidade de construção de índices
 - Verificação de restrições de integridade
 - Necessidade de ordenação dos dados
- Armazenar os dados no *data warehouse*

Armazenamento no Data Mart

